Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients

Matthieu Kowalski, Bruno Torrésani

Signal, Image and Video Processing manuscript No.
(will be inserted by the editor)

Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients

Matthieu Kowalski · Bruno Torrésani

Received: October 2007 / Revised: June 2008 / Accepted:

Abstract Sparse regression often uses ℓ_p norm priors (with $p < 2$). This paper demonstrates that the introduction of mixed-norms in such contexts allows one to go one step beyond in signal models, and promote some different, structured, forms of sparsity. It is shown that the particular case of the $\ell_{1,2}$ and $\ell_{2,1}$ norms leads to new *group shrinkage* operators. Mixed norm priors are shown to be particularly efficient in a generalized basis pursuit denoising approach, and are also used in a context of morphological component analysis. A suitable version of the Block Coordinate Relaxation algorithm is derived for the latter. The group-shrinkage operators are then modified to overcome some limitations of the mixed-norms. The proposed group shrinkage operators are tested on simulated signals in specific situations, to illustrate and compare their different behaviors. Results on real data are also used to illustrate the relevance of the approach.

Keywords Mixed-norms · Time-frequency decompositions · Sparse representations

M. Kowalski
LATP, CMI, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France
Tel.: +33-4-91054740
Fax: +33-4-91054742
E-mail: kowalski@cmi.univ-mrs.fr

B. Torrésani
LATP, CMI, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France
Tel.: +33-4-91054678
Fax: +33-4-91054742
E-mail: Bruno.Torresani@cmi.univ-mrs.fr

1 Introduction

Sparse approximation approaches have enjoyed considerable popularity in recent signal processing applications. Sparsity seems to be a particularly efficient guiding principle in view of a number of tasks such as signal compression, denoising, image de-blurring, blind source separation,... The guiding principle may be summarized as follows: for most signal classes, it is possible to find a basis or a dictionary of elementary building blocks (or atoms) with respect to which all (or most) signals in the class may be expanded, so that when the expansion is truncated in a suitable way, high precision approximations are obtained even when very few terms are retained. A large number of signal and image processing "success stories" may be described in such a way, including image compression and denoising using wavelets, curvelets, or more sophisticated *-lets, audio coding using MDCT bases, and so forth. Several efficient sparse expansion algorithms have been proposed, including among others simple expansion with respect to a fixed basis followed by soft or hard coefficient thresholding, iterative thresholding strategies in redundant dictionaries, greedy (pursuit) algorithms, or more elaborate approaches such as sparse regression in Bayesian frameworks. Thresholding and iterative thresholding strategies are particularly interesting, mainly because thresholding automatically generates sparsity. In addition, corresponding algorithms are easy to implement and generally exhibit fast convergence properties.

A main strength of these thresholding approaches is that they process the signal representation coefficient-wise, which results in low complexity algorithms. However, this may become a weakness when it comes to applications to real signals. Indeed, the assumption of

coefficient independence is generally not realistic. For example, when using wavelet or local cosine bases for expanding 1D signals, abrupt changes manifest themselves by groups of time-localized large coefficients, and frequency modulated signals exhibit *ridges* of frequency localized large coefficients. The same remark applies to edges and regular textures in wavelet or local cosine representations of images. Several different approaches have been considered to handle such dependencies between coefficients, including *structured* versions of matching pursuit (for example, harmonic or molecular versions of matching pursuit), coefficient domain modelling, or construction of suitable bases. Here, we propose to keep the coefficient modelling approach. However, rather than introducing explicit models for coefficients, we follow the thresholding and iterative thresholding approaches and design new *group thresholding* methods, associated with mixed norms in the coefficient domain.

More precisely, we consider the following problem. Let $\mathbf{y} \in \mathbb{R}^T$ be a noisy observation of a signal $\mathbf{s} \in \mathbb{R}^T$. Let $\mathcal{D}$ denote a fixed dictionary for $\mathbb{R}^T$, and denote by $A \in \mathbb{R}^{T\times N}$ the matrix whose columns are the vectors from the dictionary $\mathcal{D}$. We assume that $\mathbf{s}$ has a sparse expansion in $\mathcal{D}$, and we want to estimate $\mathbf{s}$ from $\mathbf{y}$. A classical estimate is given by the basis pursuit denoising approach introduced by Donoho and coworkers [5], also known as Tibshirani's LASSO estimate [19]. The estimate is obtained by the following optimization:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{1}$$

where $\lambda \in \mathbb{R}$ is a fixed parameter, so that, $A\hat{\mathbf{x}}$ is the estimate of $\mathbf{s}$. The ℓ_1 norm directly leads to soft thresholding strategies. Similar algorithms may be derived using more general ℓ_p norms, i.e. replacing $\|\mathbf{x}\|_1$ with $\|\mathbf{x}\|_p^p$. That estimate treats all coefficients independently. Dependencies between selected subsets of coefficients may be introduced as soon as the latter may be labelled using a double index (for example, a time-frequency index), say $\mathbf{x} = \{x_{ab}, a = 1, \dots N_a, b = 1, \dots N_b\}$. Then a new estimate is obtained by replacing the ℓ_1 norm in (1) with a mixed norm, namely by solving for

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda \left(\sum_{a=1}^{N_a} \left(\sum_{b=1}^{N_b} |x_{a,b}|^p \right)^{q/p} \right)^{1/q} . \tag{2}$$

Here, the roles of indices a and b are purely conventional. However, permuting a and b corresponds to a different problem.

The $\|.\|_{p,1}$ mixed norms have been used by various authors to model a "joint sparsity" of coefficients in the context of multichannel signals, using the FOCUSS algorithm [6], greedy pursuits [20], convex relaxation [21] or iterative thresholding strategies [12,18].

It is worth noticing that like the LASSO method and ℓ_p generalizations, the mixed norm approach admits a simple Bayesian interpretation, assuming Gaussian white noise (which justifies the choice of the ℓ_2 norm for the data fidelity term), and a coefficient prior of the form

$$f(\mathbf{x}) \;\propto\; \exp\left\{-\lambda\|\mathbf{x}\|_{p,q}^q\right\} \;,$$

which explicitly introduces couplings between coefficients.

This prior still assumes independence between groups of coefficients. We show that this independence assumption may be relaxed by modifying the design of the group-shrinkage operators used to solve (2) when A is orthogonal.

Mixed norms can also be implemented into multilayered type signal expansions, such as the ones used in [2, 8,7] for audio signals, or in the Morphological Component Analysis (MCA for short) for images [17,10]. The goal of MCA is to minimize functionals of the type

$$\Phi(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1\|_1 + \|\mathbf{x}_2\|_1 + \lambda\|\mathbf{y} - A_1\mathbf{x}_1 - A_2\mathbf{x}_2\|_2^2 \tag{3}$$

where A_1 and A_2 are the matrices corresponding to two dictionaries, chosen to be able to sparsely describe edges and textures respectively. A similar approach may be followed to separate transient and tonal layers in audio signals. According to the discussion above, we shall show that the two ℓ_1 norms in the latter expression can be conveniently replaced with suitable mixed norms, to enforce relevant dependencies between coefficients.

The paper is organized as follows. Section 2 recalls the definition of mixed norms and introduces new group-shrinkage operators associated with these norms. This section also presents the mixed norm-based multilayered expansion on union of bases. In Section 3, the shrinkage operators are modified in order to overcome some limitations of the mixed norms. All these operators are used on simulated signals to illustrate their behavior. Some applications are presented in Section 4.

2 Mixed norms and thresholding

We give in this section the definition of the mixed norms we shall be interested in. For the sake of simplicity, we shall stick to the case of two indices, even though extensions are clearly possible.

2.1 Mixed norms

We are concerned with doubly labelled sequences $x_{a,b}$, $a = 1, \dots N_a$, $b = 1, \dots N_b$. Let us start by introducing the mixed norms.

Definition 1 Let $\mathbf{x} \in \mathbb{R}^N$, labelled by a double index (a, b) . Let $p \geq 1$ and $q \geq 1$, then one can define two mixed norms $\ell_{1;p,q}$ and $\ell_{2;p,q}$ on x

$$\|\mathbf{x}\|_{1;p,q} = \left(\sum_{a=1}^{N_a} \left(\sum_{b=1}^{N_b} |x_{a,b}|^p \right)^{q/p} \right)^{1/q} , \tag{4}$$

$$\|\mathbf{x}\|_{2;p,q} = \left(\sum_{b=1}^{N_b} \left(\sum_{a=1}^{N_a} |x_{a,b}|^p \right)^{q/p} \right)^{1/q} . \tag{5}$$

The cases $p = +\infty$ and $q = +\infty$ are obtained by replacing the corresponding norm by the supremum.

Mixed norms have been used extensively by mathematicians in functional analysis (see for example [16] and references therein). Here, we limit ourselves to the finite dimensional case, and focus on the particular cases $\ell_{\bullet;1,2}$ and $\ell_{\bullet;2,1}$. For the sake of simplicity, we will use the $\ell_{1;p,q}$ norm for the theoretical study, and then denote it simply by $\ell_{p,q}$. The second case is obtained by simply switching the roles of a and b. In the numerical applications described in section 4 the choices will be specified precisely.

The reader may think of these two indices as the indices of a time-frequency signal expansion. However, let us stress that the developments below are not specific at all to time-frequency signal representations, and apply to any situation where signals are expanded with respect to a dictionary with two indices. Another simple example of that is multichannel signals, where a first index labels (scalar) dictionary elements and a second one labels channels. In an even more general situation, any discrete signal expansion may be re-labelled so as to be processed by our approach.

The two indices shall be used in hierarchical way: coefficients are split into independent groups, and coefficients within the same group are dependent. In this work, we will highlight this hierarchy by denoting the indices by g (for *group*) and m (for *member*) respectively. Using these notations, we shall label vectors $\mathbf{x} \in \mathbb{R}^N$ such that the $\ell_{p,q}$ mixed norm of $\mathbf{x}$ reads

$$\|\mathbf{x}\|_{p,q} = \left(\sum_{g=1}^{G} \left(\sum_{m=1}^{M} |x_{g,m}|^p \right)^{q/p} \right)^{1/q} ,$$

with G the number of groups and M the number of members in each group, so that $N = G \times M$.

Remark 1 Actually, nothing forces the number of members to be the same for all groups. However, by adding "phantom" members equal to zero, one can artificially come back to the simplest situation where all groups have the same size. For the sake of simplicity, we only consider that simple case here.

It is interesting to stress that a $\ell_{p,q}$ mixed norm can be seen as a "composition" of ℓ_p and ℓ_q norms, and therefore inherits of their properties (in particular convexity for $p, q \geq 1$). With the above notations,

$$\|\mathbf{x}\|_{p,q} = \left(\sum_{g=1}^{G} \|\mathbf{x}_g\|_p^q \right)^{1/q} = \|(\|\mathbf{x}_1\|_p, \dots, \|\mathbf{x}_G\|_p)\|_q . \tag{6}$$

For $p < 2$, ℓ_p norms are often used as *diversity* measures, and minimizing the ℓ_p norm of a coefficient sequence of a signal generally aims at promoting *concentration* for the expansion: the distribution of coefficients is more sharply peaked at the origin for $p < 2$. For $p \leq 1$, concentration becomes sparsity, since small coefficients are forced to zero. The case $p = 1$ has a particular status, since the ℓ_1 norm promotes sparsity and remains convex. The situation with mixed norms is a bit more tricky, since two exponents have to be taken into account. However, we shall see below that values of p (or q) smaller than 2 still yield some form of concentration, in a somewhat *structured* way. More precisely, depending on the choice of p and q, concentration is promoted on each individual variable $x_{g,m}$ if p is close to 1, and on an entire group of variables if q is close to 1.

2.2 Group-shrinkage operators

We first introduce generalized shrinkage operators, extending LASSO and Group-LASSO (G-LASSO) estimators, before turning to extensions to the multilayered case. For the sake of simplicity, we shall concentrate on the particular values $(p, q) \in \{1, 2\}$.

Given an observation $\mathbf{y} \in \mathbb{R}^N$, let us denote, for $\mathbf{x} \in \mathbb{R}^N$,

$$\Phi_{p,q}[\mathbf{x}] = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\lambda}{q}\|\mathbf{x}\|_{p,q}^q , \tag{7}$$

We want to solve the following optimisation problem

$$\mathcal{P}_{p,q} : \qquad \hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{R}^N}{\operatorname{argmin}} \Phi_{p,q}[\mathbf{x}]$$

in the particular case where A is an orthogonal matrix. For the sake of simplicity, let us introduce the following notation

$$\bar{\mathbf{y}} = A^T \mathbf{y} . \tag{8}$$

Then problem $\mathcal{P}_{pq}$ can also be written

$$\min_{\mathbf{x}\in\mathbb{R}^N} \left[\frac{1}{2}\|\bar{\mathbf{y}} - \mathbf{x}\|_2^2 + \frac{\lambda}{q} \sum_{g=1}^{G} \left(\sum_{m=1}^{M} |x_{g,m}|^p \right)^{q/p} \right] , \qquad (9)$$

Focusing on the case $p, q \in \{1, 2\}$, and denoting by g and m the indices as explained before (i.e. $\mathbf{x} = \{x_{g,m}\}$), we focus on problems $\mathcal{P}_{1,2}$ and $\mathcal{P}_{2,1}$.

The solution is given in Proposition 1 below. Let us first introduce some notations. For $x \in \mathbb{R}$, we shall set $x^+ = x$ if $x \geq 0$ and $x^+ = 0$ if $x \leq 0$. For $\tau \in \mathbb{R}^+$, we denote by $\mathbb{S}_\tau : \mathbb{R} \to \mathbb{R}$ the soft thresholding operator

$$\mathbb{S}_\tau(x) = \begin{cases} \mathrm{sgn}(x)(|x| - \tau) & \text{if } |x| \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

Also, given a vector $\bar{\mathbf{y}}_g = \{\bar{y}_{g,1}, \dots \bar{y}_{g,M}\}$, denote by $\check{\mathbf{y}}_g = \{\check{y}_{g,1}, \dots \check{y}_{g,M}\}$ the vector whose components are the absolute values of coefficients $\bar{y}_{g,m}$, sorted in descending order: $\check{y}_{g,1} \geq \check{y}_{g,2} \geq \cdots \geq \check{y}_{g,M}$. Finally, for a given $\lambda \in \mathbb{R}^+$, denote by $M_g(\lambda)$ the positive integer such that

$$\check{y}_{g,M_g(\lambda)+1} \leq \sum_{m=1}^{M_g(\lambda)+1} (\check{y}_{g,m} - \check{y}_{g,M_g(\lambda)+1})$$

and

$$\check{y}_{g,M_g(\lambda)} > \lambda \sum_{m=1}^{M_g(\lambda)} (\check{y}_{g,m} - \check{y}_{g,M_g(\lambda)}) ,$$

and set

$$\|\bar{\mathbf{y}}_g\| = \sum_{m=1}^{M_g(\lambda)} \check{y}_{g,m} = \|\check{\mathbf{y}}_{g,1:M_g(\lambda)}\|_1 ,$$

where $\check{\mathbf{y}}_{g,1:M_g(\lambda)}$ denotes the vector $\{\check{y}_{g,1}, \dots \check{y}_{g,M_g(\lambda)}\}$.

Proposition 1 *Let A be an orthogonal matrix.*

(a) The solution $\hat{\mathbf{x}}$ of problem $\mathcal{P}_{1,2}$ is given by the following shrinkage operation: for all g, m

$$\hat{x}_{g,m} = \mathbb{S}_{\tau_g}(\bar{y}_{g,m}) ,$$

where the group dependent threshold τ_g reads

$$\tau_g = \frac{\lambda}{1 + \lambda M_g(\lambda)} \|\bar{\mathbf{y}}_g\|$$

(b) The solution $\hat{\mathbf{x}}$ of problem $\mathcal{P}_{2,1}$ is given by the following shrinkage operation: for all g, m

$$\hat{x}_{g,m} = \bar{y}_{g,m} \left(1 - \frac{\lambda}{\|\bar{\mathbf{y}}_g\|_2} \right)^+ .$$

Remark 2 The solution of $\mathcal{P}_{2,1}$ is known in the statistical community as the G-LASSO estimate, and the result was given in [23]. The solution of problem $\mathcal{P}_{1,2}$ is obtained in [14] as a part of a more general result. In contrast with the G-LASSO, we call the problem $\mathcal{P}_{1,2}$ the Elitist-LASSO (E-LASSO, see below). Notice that in both cases, the result is a generalized soft thresholding, or shrinkage, that is applied to a group of coefficients rather than single coefficients. Hence, coefficients are not processed independently any more.

It is important to stress the striking difference between the two new shrinkage operators. In the second case (the G-LASSO case), a 1D group of coefficients is either globally retained or discarded. This may be understood as an *united group shrinkage*, since the same threshold applies to all members of a given group. In the first case, each coefficient is shrunk individually, but the corresponding threshold depends on its 1D neighborhood. That one can be understood as an *elitist group shrinkage*, since most members of a given group are thresholded, and only the emerging coefficients of each group (it may be shown that at least one coefficient is kept, see [14]) remain. The difference between these two situations will appear clearly in the numerical results below. There, we also present an approximate solution of $\mathcal{P}_{1,2}$, which turns out to be computationally simpler.

It is also interesting to remark that the solution of $\mathcal{P}_{1,2}$, which only involves soft thresholdings with variable threshold values, is also the solution of the problem

$$\min_{\mathbf{x}\in\mathbb{R}^N} \left[\sum_{g=1}^{G} \frac{1}{\tau_g} \sum_{m=1}^{M} |y_{g,m} - [A\mathbf{x}]_{g,m}|^2 \; + \; \lambda\|\mathbf{x}\|_1 \right] , \qquad (10)$$

i.e. a sparse regression problem, with ℓ_1 sparsity prior, and in which the data fidelity term involves a data dependent weighting. From a Bayesian point of view, such a re-interpretation shows that beyond the Gaussian white noise case, the so obtained solution may also be expected to perform well in situations where a sparse signal is embedded into a noise whose variance varies as a function of g in the coefficient domain.

2.3 Multilayered expansion on union of bases

After having found the solution of problem $\mathcal{P}_{pq}$ in the simple case where A is an orthogonal matrix (corresponding to an orthonormal basis), we now address similar problems, in which A is a concatenation of orthogonal matrices (corresponding to an union of orthonormal bases), and the coefficient priors are different for each basis. A motivation for this problem is the decomposition of audio signals into three layers *Transient + Tonal + Noise*, using MDCT bases

with different time-frequency resolutions. Similar problems may also be found in image processing, under the name of *Cartoon + Texture + Noise* image decompositions. Such problems have been studied by various authors and a few algorithms are already available. Probability-based approaches have been used in the audio domain (see [15,11], that exploit simultaneously sparsity and persistence. Variational approaches (such as the so-called Morphological Component Analysis) were generally preferred in the image processing literature, that did not so far integrate the notion of persistence. The mixed norm approach we focus on represents a good compromise between the other two approaches, as it allows one to incorporate persistence in variational formulations.

We start from an optimization problem similar to the one given by MCA, but, instead of using two ℓ_1 norms to estimate the tonal and transient layers, we will use suitable mixed-norms. So that, we will minimize the following functional

$$\Phi(\mathbf{x},\tilde{\mathbf{x}}) = \|\mathbf{y} - A(\mathbf{x},\tilde{\mathbf{x}})^T\|_2^2 + \lambda\|\mathbf{x}\|_{p,q}^q + \mu\|\tilde{\mathbf{x}}\|_{\tilde{p},\tilde{q}}^{\tilde{q}} \qquad (11)$$

where the $\ell_{p,q}$ and $\ell_{\tilde{p},\tilde{q}}$ norms will be chosen adequately.

To decompose a signal into several layers, one chooses a suitable dictionary for each layer. In the audio signal example, the transient layer is known to be sparsely represented in dictionaries of wavelets, or time-frequency dictionaries (like Gabor or MDCT) with a narrow window. At the opposite, the tonal layer is known to be sparsely represented in time-frequency dictionaries with a wide window.

Here we choose the special case where each dictionary is an orthonormal basis, for example, two MDCT bases with two different sizes for the windows, and apply the Block Coordinate Relaxation method [4] (BCR for short) which inspired the Morphological Component Analysis (MCA) algorithms [17]. BCR is specially adapted to unions of orthogonal bases, and is known to converge to a minimum of the *basis-pursuit denoising* objective functional (3).

Let us introduce some notations. We denote by $\mathbf{U}$ and $\mathbf{V}$ the two bases under consideration, and by U and V the corresponding matrices. We denote by $\mathbf{x_U}$ the coefficients corresponding to the basis $\mathbf{U}$ and $\mathbf{x_V}$ the coefficients corresponding to the basis $\mathbf{V}$. So that, $U\mathbf{x_U}$ corresponds to the tonal layer and $V\mathbf{x_V}$ to the transient layer. To obtain estimates for the two layers, we then choose to minimize the following functional

$$\Phi(\mathbf{x_U},\mathbf{x_V}) = \frac{1}{2}\|\mathbf{y}-U\mathbf{x_U}-V\mathbf{x_V}\|_2^2+\frac{\lambda}{q}\|\mathbf{x_U}\|_{p,q}^q+\frac{\mu}{\tilde{q}}\|\mathbf{x_V}\|_{\tilde{p},\tilde{q}}^{\tilde{q}} \qquad (12)$$

The BCR algorithm is then slightly modified in order to yield a minimizer of (12):

Algorithm 1

- **Let** $\mathbf{x}_\mathbf{U}^{(0)} \in \mathbb{R}^N$ *and* $\mathbf{x}_\mathbf{V}^{(0)} \in \mathbb{R}^N$
- **Do**
  1. $\mathbf{r}_\mathbf{U}^{(i)} = \mathbf{y} - V\mathbf{x}_\mathbf{V}^{(i)}$
  2. *Find an estimate* $\mathbf{x}_\mathbf{U}^{(i+1)}$ *by solving*

$$\mathbf{x}_\mathbf{U}^{(i+1)} = \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - U\mathbf{x}\|_2^2 + \frac{\lambda}{q}\|\mathbf{x}\|_{p,q}^q$$

  using Proposition 1
  3. $\mathbf{r}_\mathbf{V}^{(i)} = \mathbf{y} - U\mathbf{x}_\mathbf{U}^{(i+1)}$
  4. *Find an estimate* $\mathbf{x}_\mathbf{V}^{(i+1)}$ *by solving*

$$\mathbf{x}_\mathbf{V}^{(i+1)} = \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - V\mathbf{x}\|_2^2 + \frac{\mu}{\tilde{q}}\|\mathbf{x}\|_{\tilde{p},\tilde{q}}^{\tilde{q}}$$

  using Proposition 1
- **Until** *convergence*

Following the proof given in [4] for the BCR algorithm, one can exploit the results of [22] and state

Theorem 1 *Let $U, V \in \mathbb{R}^{N\times N}$ be two orthogonal matrices. Let $\mathbf{y} \in \mathbb{R}^N$ and $p \geq 1$, $q \geq 1$ $\tilde{p} \geq 1$, and $\tilde{q} \geq 1$. Then Algorithm 1 converges to a minimum of* (12).

Remark 3 Here, we do not deal with the more general case where the dictionary is an arbitrary frame or even an union of frames. This case was studied in [14], where convergence of a corresponding iterative thresholded Landweber algorithm was proven. However, let us point out that the approach used in the MCA algorithm could also be used if $\mathbf{U}$ and $\mathbf{V}$ are two frames. The heuristics of this algorithm is to decrease the parameters λ and μ in (3) during the iterations. Numerical results seem to indicate convergence to the global minimum, though no formal proof has been given so far (to our knowledge). This heuristics was then generalized by the Stagewise Matching Pursuit [9].

3 Group shrinkage in practice: simulations

The main goal of this section is to illustrate and comment on the choice of the shrinkage operator, for specific problems on simulated signals. To this end, we limit ourselves to decompositions on an orthogonal basis (i.e. A is an orthogonal matrix). We introduced in the previous section two particular generalized shrinkage operators (G-LASSO and E-LASSO), with two completely different behaviors in an orthogonal basis. Here we analyze and illustrate the behaviors of these two approaches, and propose alternatives that overcome some potential shortcomings in specific situations.

To this end, we applied G-LASSO and E-LASSO and variants to simulated signals, specifically designed to illustrate their behavior. The simulated signals were obtained as follows. First, the time-frequency map Λ was simulated, using the Hidden Markov Model studied in [15], that generates persistence along the time axis. This map was then used to simulate a signal of the form

$$y[t] = \sum_{\ell \in \Lambda} x_\ell u_\ell[t] \ . \tag{13}$$

The index set $\Lambda = \{\ell : x_\ell \neq 0\}$ is called the *significance map* or *time-frequency map*. The corresponding (i.i.d) time-frequency coefficients x_ℓ were simulated using a normal law $\mathcal{N}(0,1)$.

An example of so-generated significance map is displayed in Fig. 1. The map has 8.5% non zero coefficients. In the numerical examples shown below, we will consider maps with 8.5% and 1% non zero coefficients respectively, to study the behavior of the algorithms at different sparsity levels.

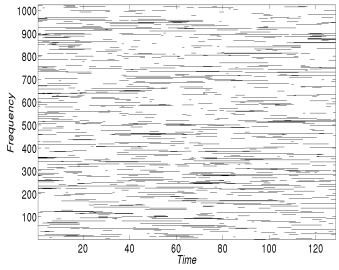

Fig. 1 Time-Frequency map with 8.5% non-zero coefficients, generated using fixed frequency Markov chains.

3.1 Selection of relevant groups

3.1.1 Relabelling

In some situations, all the members of a given group need not be active at the same time. When subgroups that are simultaneously active or inactive are known in advance, coefficients may be re-labelled so that the classical G-LASSO estimate may be used. A trivial example of such re-labelling is shown in Fig. 2, in the context of a multichannel signal. There, the re-labelling is simply a splitting of groups into subgroups, but more complex re-labellings can also be considered.

However, this is not the most general situation, and subgroups of active coefficients are generally neither known in advance, nor even fixed. For that reason, a "sliding window" alternative of the above described approach is desirable.

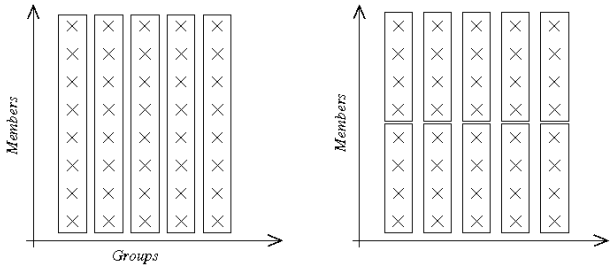

Fig. 2 An example of coefficient re-labelling.

3.1.2 Windowing

For this purpose, let us now assume that some extra information about the coefficients is available, telling us for each coefficient of index $k = (g, m)$ which are the other coefficients that are likely to be "active" or "inactive" simultaneously with k. This generates a neighborhood system, associating to any "group-member" index $k = 1, \dots N$ a group $\mathcal{N}(k)$ of "close" indices. Now, for a given index k, it seems reasonable to use only its neighbors in $\mathcal{N}(k)$ to estimate its sparse expansion, exploiting persistence within $\mathcal{N}(k)$. Using the G-LASSO estimate in Proposition 1-(b), this suggests to compute

$$\hat{x}_{g,m} = \bar{y}_{g,m} \left(1 - \frac{\lambda}{\|\bar{\mathrm{y}}_k\|_{\ell_2(\mathcal{N}(k))}}\right)^+ \ , \tag{14}$$

where we have denoted by $\bar{\mathrm{y}}_k$ the subsequence

$$\bar{\mathrm{y}}_k = \{\bar{y}_{k'}, k' \in \mathcal{N}(k)\} \ .$$

We call this estimate the Windowed Group-LASSO (WG-LASSO). Notice that unlike the re-labelling approach alluded to above, each coefficient k uses its own neighbors, instead of the whole group. Compared to Proposition 1-(b), the estimated coefficients $\hat{x}_{g,m}$ are obtained from the observations $\bar{y}_{g,m} = [Ay]_{g,m}$ by pointwise multiplication with a mask function, which now depends on the index $k = (g, m)$. Notice also that this new generalized thresholding is not any more associated to a simple variational problem. Fig. 3 shows an example of a sliding window used to group a channel with its neighborhood, where m is the channel index and g the time-frequency index.

3.1.3 Simulations

Here we consider decomposition of multichannel signals (sampled at 44100 Hz) on a given MDCT basis (with 23.3 millisecond long window):

$$y_m[t] = \sum_{m \in \Lambda} x_{g,m} u_g[t] \ , \tag{15}$$

where g is a time-frequency index and m labels channels. $\Lambda = \{g : x_{g,m} \neq 0\}$ is the *significance map*, or

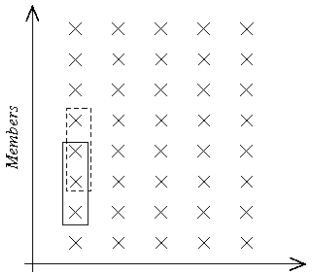

Fig. 3 An example of sliding window.

time-frequency map, and is assumed to be the same for all channels. Then we denote by $\mathbf{y} = (\mathbf{y}^1, \dots, \mathbf{y}^M) \in \mathbb{R}^{N\times M}$ the multichannel signal, organized as an $N \times M$ matrix whose columns are the channels, and $\mathbf{x} \in \mathbb{R}^{N\times M}$ the unknown coefficient sequences. As before, we set $\overline{\mathbf{y}} = A\mathbf{y} \in \mathbb{R}^{N\times M}$ (we recall that here A is an $N \times N$ orthogonal matrix).

In this context, the groups labelled by g and the members labelled by m in the previous section correspond respectively to the time-frequency indices and the channels. In other words, the model involves "between channels" dependencies.

Two multichannel signals were simulated as follows

1. Choose a percentage of non-zero coefficients, and generate two time frequency maps Λ_1 and Λ_2 with that prescribed percentage.
2. Simulate two sets of i.i.d. $\mathcal{N}(0, 1)$ time-frequency coefficients $x_{g,m}$, $m = 1, \dots 4$ and $g \in \Lambda_1$ (resp. $m = 5, \dots 8$ and $g \in \Lambda_2$).
3. Synthesize the signals using model (15).

The simulated signals have then $M = 8$ channels. The first four channels share time-frequency map Λ_1 and the last four share time-frequency map Λ_2.

The various generalized thresholding estimators described above are compared in the context of a denoising problem. A Gaussian white noise is added to the multichannel signals so as to obtain a SNR equal to 10 dB. For the sake of simplicity, the SNR is not calculated channelwise, but on the the entire multichannel signal:

$$SNR(\mathbf{x}, \hat{\mathbf{x}}) = 20 \log_{10} \left(\frac{\|\mathbf{x}\|_2}{\|\mathbf{x} - \hat{\mathbf{x}}\|_2} \right) \tag{16}$$

where $\|.\|_2$ denotes the Fröbenius norm of the multichannel signal. This SNR may differ from the mean of SNR of all the channel, but this difference is less than 1 dB and does not influence the behavior of the displayed curves.

The estimators under study are the following

– LASSO, corresponding to the problem

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^{N\times M}}{\operatorname{argmin}} \|\bar{\mathbf{y}} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \ .$$

All "channel-time-frequency" coefficients are independent, the estimate is obtained by soft-thresholding.

– G-LASSO 1, corresponding to the problem

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^{N\times M}}{\operatorname{argmin}} \|\bar{\mathbf{y}} - \mathbf{x}\|_2^2 + \lambda \sum_{g=1}^{G} \left(\sum_{m=1}^{M} |x_{g,m}| \right)^2 .$$

For a given time-frequency index, all the channels are gathered to create the groups of G-LASSO. The groups are independent. This corresponds of the grouping given on Fig. 2 (left).

– G-LASSO 2, which exploits prior information on the two time-frequency maps Λ_1 and Λ_2, corresponds to

$$\min_{\mathbf{x}\in\mathbb{R}^{N\times M}} \|\bar{\mathbf{y}} - \mathbf{x}\|_2^2 + \lambda \sum_{g=1}^{G} \left[\left(\sum_{m_1=1}^{4} |x_{g,m_1}| \right)^2 + \left(\sum_{m_2=5}^{8} |x_{g,m_2}| \right)^2 \right] .$$

For a given time-frequency index, the first four channels are gathered into a group, and the last four into an another group. The groups are independent and correspond of the regrouping given on Fig. 2 (right).

– The WG-LASSO, corresponding to the estimate given in Equation (14). The two nearest neighbors of a channel are gathered using a sliding window to give the estimate. This corresponds to the grouping given in Fig.3.

Different estimates were computed for various values of λ. The range of values for λ was chosen so as to obtain estimates with different degrees of sparsity, i.e. with various numbers of coefficients set to zero: the bigger the λ, the sparser the estimate.

The curves in Fig. 4 show the evolution of SNR as a function of the number of non zero coefficients (which depends on the value of λ) of the different estimates for the simulated signals using the map displayed in Fig. 1 (i.e. 8.5% nonzero coefficients). Similar results, obtained using sparser significance maps (1% nonzero coefficients) are displayed in Fig. 5.

The behavior of the estimates clearly depends on the degree of sparsity of the input signal. In the two considered cases, G-LASSO 2 (which uses more prior information than the others) reaches the best SNR, and provides an higher SNR than the other estimates when the number of selected coefficients is close to the true number of non zero coefficients.

WG-LASSO outperforms LASSO on the two curves, except for large values of the sparsity penalty (when many coefficients are set to zero); however this case must be avoided to obtain a good estimate, as the SNR collapses quickly.

Despite their globally different aspects, the curves in Fig. 4 and Fig. 5 show very similar behaviors. In

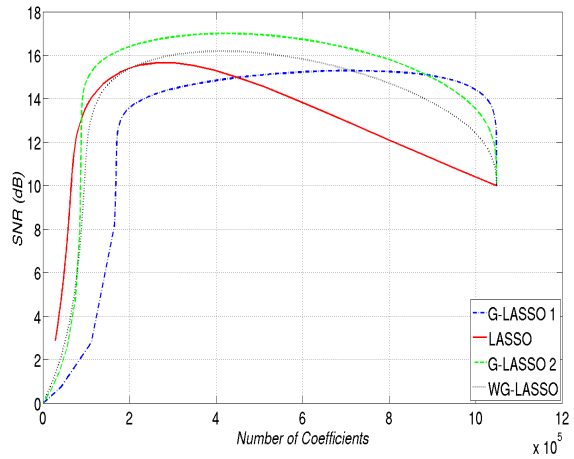

Fig. 4 Comparison between LASSO, 2 types of G-LASSO and WG-GLASSO.

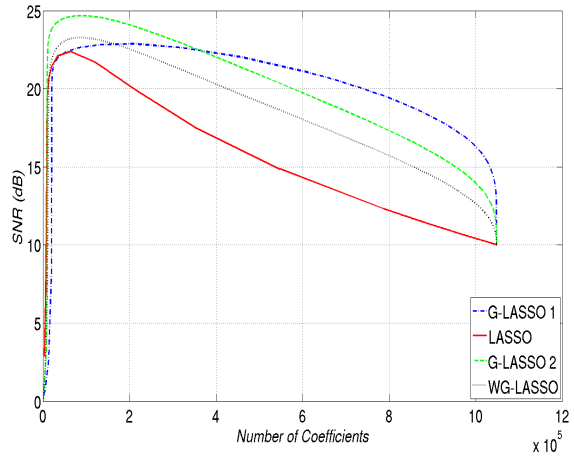

Fig. 5 Comparison between LASSO, 2 types of G-LASSO and WG-GLASSO.

both cases, best results are obtained when the groupings are known *a priori*. When such a prior information is unavailable, the WG-LASSO is definitely a good alternative to exploit dependences between some coefficients. In addition, if coefficients cannot be clustered into groups, but possess some neighboring relationships, WG-LASSO is able to exploit the latter.

3.2 Coefficient selection within sparse groups

Let us now turn to E-LASSO estimates. We show in this subsection some numerical results obtained using E-LASSO, a simplified version of E-LASSO, and a variant proposed in order to introduce *across groups persistence*.

3.2.1 An approximation of the E-LASSO estimate

As may be seen from Proposition 1, E-LASSO involves a sorting of coefficients $\bar{y}$, and the determination of numbers $M_g(\lambda)$ prior to the actual shrinkage operation. When these operations are skipped, this yields an approximation of the estimator, obtained by replacing the threshold

$$\tau_g = \frac{\lambda}{1+\lambda M_g(\lambda)} |||\bar{\mathbf{y}}_g|||$$

by the approximation

$$\tau'_g = \frac{\lambda}{1+\lambda M} \|\bar{\mathbf{y}}_g\|_1 \ . \tag{17}$$

This approximation called AE-LASSO (for Approximate E-LASSO), is simpler to compute, and has a practical interpretation, in particular in the limit of large λ values. Indeed, letting $\lambda \to \infty$ in Equation (17), the coefficient are thresholded by $\|\bar{y}_{g,m}\|_1/M$, which is the average of the coefficients $|\bar{y}_{g,m}|$ for a fixed group index g. The main shortcoming of this approximation is that the threshold is bounded by this average value, which bounds from below the number of retained coefficients. An advantage of this approximation, is that the role of the regularization parameter is much easier to understand. We shall see in the numerical examples below that when the number of retained coefficients is large enough, AE-LASSO is actually a good approximation of E-LASSO.

3.2.2 Introduction of persistence

As we have seen above, the $\ell_{1,2}$ coefficient penalty is significantly different from the $\ell_{2,1}$ one, that leads to G-LASSO regression: it promotes sparsity within groups of coefficients instead of sparsity across groups. For example, the thresholding formula (17) selects a small number of coefficients within each group. To fix the ideas, let us assume that a single coefficient is retained within each group. This coefficient is likely to vary from a group to another, since nothing in the norm prevents it from doing so. If one wants to promote persistence in the retained coefficients, an approach similar to the previous one may be developed, taking into account neighbors of the considered coefficients. We start by associating to any group index g a family $\mathcal{N}(g)$ of neighbors. Then, for fixed g, we can solve the minimization problem with $\ell_{1,2}$ coefficient penalty on the vector $\bar{\mathbf{y}}_{\mathcal{N}(g)} = \{\bar{\mathbf{y}}_{g',m}, m = 1, \dots M, g' \in \mathcal{N}(g)\}$

Applying the same approach as before, the generalized thresholding formula (17) is now replaced with

$$\tau''_g = \frac{\lambda}{1+\lambda|\mathcal{N}(g)|} \|\bar{\mathbf{y}}_{\mathcal{N}(g)}\|_1 \ , \tag{18}$$

$|\mathcal{N}(g)|$ being the cardinality of the set $\mathcal{N}(g)$. Again, this generalized thresholding is not associated with a simple variational approach. The corresponding estimator is termed PE-LASSO (for Persistent Elitist LASSO).

3.2.3 Simulations

To illustrate the behavior of the estimators described above, we simulated a signal as follows. First, a time-frequency map was generated as before (the map 1 of the previous subsection was chosen); then coefficients were generated from a normal law $\mathcal{N}(0,1)$. To follow the model given by Equation (10), at each time index, we added a Gaussian white noise, whose variance was randomly taken from a uniform distribution (between 1 and 128). Then denoising was performed using the E-LASSO, AE-LASSO, PE-LASSO, WG-LASSO and LASSO estimates, with various values of the λ parameter. The PE-LASSO estimate was done by introducing time persistence, taking nearest neighbors into account (1 time index before and one after). WG-LASSO was performed by gathering the 4 time-neighbors of a given time-frequency coefficient.

We display in Fig. 6 and 7 the SNR as a function of the number of retained non-zero coefficients for the previous estimators, for two different values of input SNR. As expected, the E-LASSO estimate performs best in this situation. The AE-LASSO estimate is close of the E-LASSO estimate, but does not allow for very small numbers of retained coefficients (as explained before). WG-LASSO performs quite well when the number of non zero coefficients is over-estimated. Finally, the PE-LASSO estimate is quite disappointing, as it only outperforms the classical LASSO. Introducing persistence into the estimator does not seem to pay, even in situations where persistence is present in the signal.

4 Results on real signals

We now illustrate the various approaches described above with three different problems:

- Denoising of multichannel signals, in an additive Gaussian white noise situation.
- Denoising of a single channel signal, with non stationary random noise
- Multilayered signal decomposition.

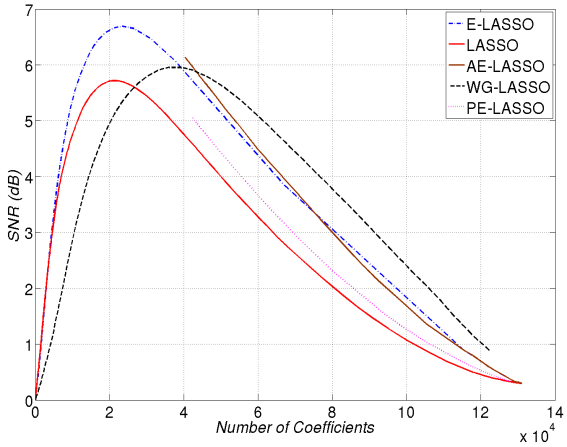

Fig. 6 Comparison between LASSO, E-LASSO, AE-LASSO (approximation of E-LASSO), PE-LASSO (E-LASSO with persistence) and WG-LASSO; input SNR=3dB.

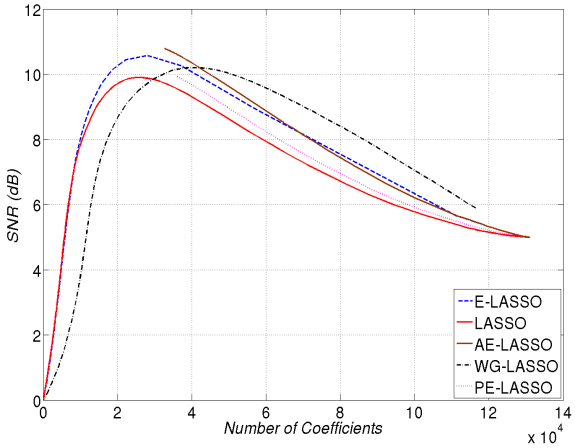

Fig. 7 Comparison between LASSO, E-LASSO, AE-LASSO, PE-LASSO and WG-LASSO; input SNR=5dB.

4.1 Multichannel denoising

Sparse approximation techniques have been extended recently to multichannel signals (see [3,13] and references therein). We address such a problem directly via a generalized basis pursuit denoising approach, using the ℓ_1 norm in the time-frequency direction, and the ℓ_2 norm across channels.

Let us consider a multichannel signal $\mathbf{y} = \{y_{gm},\ g = 1,\dots G,\ m = 1,\dots M\}$, g denoting the time index and M the channel index. Consider an orthonormal basis $\mathbf{U} = \{u_g, g = 1,\dots G\}$ (here, g labels the atoms of the basis) for the single channel signal space. We are interested in expansions of the form $\mathbf{y} = \sum_g \mathbf{x}_g u_g$ (where multichannel vectors are denoted with bold symbols), in cases where the observations are noisy, and the basis $\mathbf{U}$ has been chosen in such a way that the coefficient

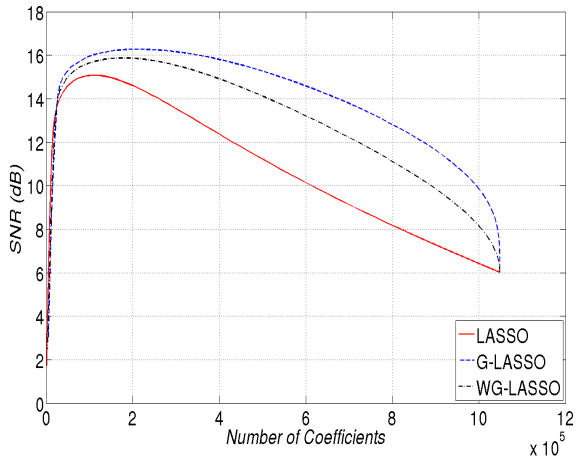

Fig. 8 Multichannel denoising of the train signal. SNR as a function of the number of retained coefficients; full curve: LASSO; dashed curve: G-LASSO; dashed-dotted curve: WG-LASSO.

sequences $\mathbf{x}$ are sparse in the g direction, and persistent across channels. Then, we are close to the case described in Section 3.1.

In this case, A is an orthogonal matrix and the optimization problem is formulated as before:

$$\min_{\mathbf{x}\in\mathbb{R}^{N\times M}} \left(\|\mathbf{y}-A\mathbf{x}\|_2^2+\lambda\|\mathbf{x}\|_{p,q}^q\right) \ ,$$

M being the number of channels. Since the matrix A remains orthogonal, the results above may then be applied directly.

Since we aim at privileging groups of coefficients (persistence across channels), we choose the G-LASSO estimate provided in Proposition 1-(b). We illustrate this problem with a sound example recorded in a running train.

The considered signal features low frequency noise, phone ringings, voice, clicks and additional transient components. The signal is a four channels signal, recorded using three directional and one omni-directional microphones. Gaussian white noise was added to the four channels, yielding input SNR equal to 6 dB. The signal was denoised by applying LASSO (corresponding to the ℓ_1 norm prior on the set of coefficients), and G-LASSO (corresponding to $\ell_{2,1}$ norm prior on coefficients). As stressed before, this choice is motivated by the desire of using the same significance map (i.e. the set of labels of nonzero coefficients) for all channels. Simulations were run with various values of the threshold (i.e. the Lagrange parameter). Corresponding SNR curves are displayed in Fig. 8.

The mixed norm based approach clearly outperforms the ℓ_1 norm approach significantly. Similar results (not shown here) were also obtained on different multichannel audio signals. The improvement appears to increase with the number of channels, as may be expected. In the particular example considered here, we remark that even though the four microphones are different (three being directional), the four signals are coherent enough for G-LASSO to improve significantly the LASSO results.

Let us finally stress that the same approach may be developed in many other multichannel signal denoising contexts, such as color image denoising, multispectral imaging,...

4.2 Denoising a "vinyl recording" like noisy signal

The E-LASSO and AE-LASSO estimates are now compared to the LASSO in the context of single channel denoising. In the standard *additive Gaussian white noise* benchmark, the soft-thresholding provided by LASSO is a better choice than the generalized shrinkage operators obtained using mixed-norms. However, based on Equation (10), we also remark that E-LASSO and AE-LASSO are valuable alternatives when going beyond the Gaussian white noise assumption, in cases where the noise variance varies with the group index g (this was already visible in the experiments of Section 3.2).

Here we consider the case of (single channel) audio signal, perturbed by additive non-stationary noise, whose variance varies significantly with time. The considered example was taken from vinyl recordings[1]. Vinyl recording noise (including many "cracks" and other non-stationary noises) was added to a musical signal (excerpt of about 6 s, 2^{18} samples at 44100 Hz sampling rate, of the song "Mamavatu" from Susheela Raman), the resulting input SNR being about 1 dB only. This noisy signal was then expanded in a MDCT basis (with 512 samples -about 11 ms- long windows). The group index g (see Section 2) was chosen as the time index of the MDCT basis functions, and m as the frequency index.

Fig. 9 displays the evolution of the output SNR as a function of the number of non-zero coefficients for the LASSO, E-LASSO, AE-LASSO, PE-LASSO and WG-LASSO. The behavior of the curves is close to the one observed on Fig. 6 and 7 of Section 3.2, even though the SNR improvement is not large.

4.3 Multilayered audio signal expansion

We now consider the problem of decomposing (single channel) signals into layers of different nature, focusing again on the case of audio signals, from which we

[1] Samples of vinyl recordings noise are available at the web site `www.universal_soundbank.com/audio.htm`

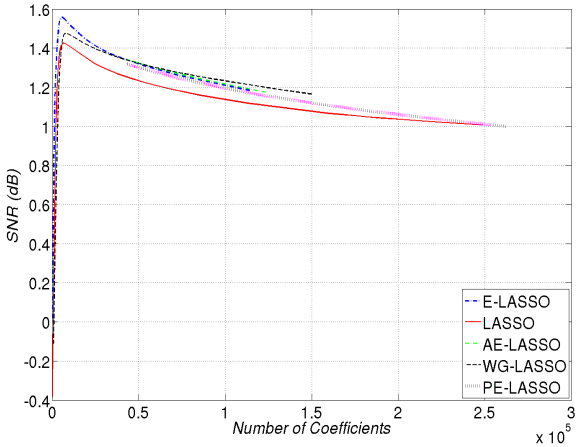

Fig. 9 Comparison between LASSO, E-LASSO, AE-LASSO, PE-LASSO and WG-LASSO, on the single channel audio signal with additive "vinyl recording type" noise.

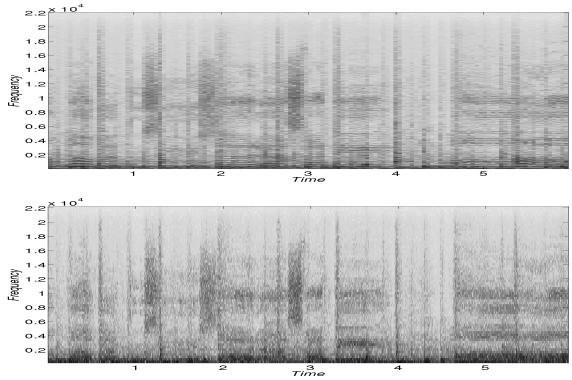

Fig. 10 MDCT coefficients of th signal. Top: in $\mathbf{U}$, bottom: in $\mathbf{V}$.

aim at extracting transient and tonal components. This problem has received increasing interest recently, as such a separation proves useful both by itself (for example for denoising [11] or compression [15], and as pre-processing step for various sound processing tasks (such as, signal analysis, for which different layers are analyzed using different approaches, or sound transformation, e.g. pitch shifting, for which the transformation has to be different for different layers). An intrinsic difficulty is the absence of ground truth that could be used for validating the proposed approaches. However, we shall see that the approach developed here is suitable for multilayered signal decomposition, and that different choices for the coefficient priors yield significantly different results.

Multilayered separation may be performed using various approaches [11,15] (see also the MCA algorithm for cartoon + texture separation in images [17]). Here we illustrate the influence of the mixed-norm in the regression problem (12), in comparison to the usual ℓ_1 norm used in the MCA regression problem (3).

We choose a musical signal taken from the "Mamavatu" song (see above), that involves percussive instruments, voice and guitar. The signal duration is about 6 s (2^{18} samples). Keeping the notations of subsection 2.3, one then expects to obtain an estimate $V\hat{x}_{\mathbf{V}}$ of the transient layer, and an estimate $U\hat{x}_{\mathbf{U}}$ of the tonal layer. We compare the estimates given by choosing two ℓ_1 norms (as in MCA), and several mixed norms, to be specified below. We choose for $\mathbf{U}$ a MDCT basis with a 4096 samples window length, and for $\mathbf{V}$ a MDCT basis with a 128 samples window length. The representations of the MDCT coefficients of the tonal (resp. transient) layer in $\mathbf{U}$ (res. $\mathbf{V}$) are shown in Fig. 10. The particular structures of both layers, with their persistence properties, appear clearly there.

The tonal layer is expected to be sparsely represented in the frequency domain, with emergent frequencies that may evolve slowly with time (i.e. almost horizontal lines of large MDCT coefficients). Possible choices for the estimates are E-LASSO (sparse within group) with the time label as group label, or G-LASSO (sparse across groups) with the frequency label as group label. However, the latter choice turns out to be a poor strategy for the tonal layer, because of the slow evolution in time of the frequencies. Furthermore, experiments show that for this example, LASSO and E-LASSO performs very similarly to estimate the tonal layer. We thus limit the present illustration to LASSO estimates for the tonal layer.

In a similar spirit, the transient layer is expected to be sparse in time, but spread out in the frequency domain. Then, for the transient layer, while E-LASSO with the frequency label as group label still seems a relevant choice, G-LASSO (with the time label as group label) is also interesting because of the particular structures of transients, which are most often sharply time-localized.

Then, in order to show the differences between the ℓ_1, ℓ_{12} and ℓ_{21} mixed-norms, we fixed the ℓ_1 norm to estimate the tonal layer, and we compared the results with the three different choices for the transient layer.

In the numerical simulations presented here, the λ and μ parameters were tuned to obtain approximately the same number of coefficients for each functional, to well illustrate the differences between the norms. Table 1 summarizes the results obtained using the three possible functionals. The first line of the table gives the choices that were made for norms for the tonal layer and the transient layer. The second and the third (resp fourth and fifth) lines give respectively the numbers of retained coefficients for $x_{\mathbf{U}}$ (resp $x_{\mathbf{V}}$) and SNR of this

norms	L / L	L / EL	L / GL
nbcoeff $x_{\mathbf{U}}$	16.4%	16.5%	16.7%
SNR $x_{\mathbf{U}}$	7.6 dB	17.8 dB	20.2 dB
nbcoeff $x_{\mathbf{V}}$	7.4%	7.5%	7.4%
SNR $x_{\mathbf{V}}$	2.8 dB	0.22 dB	0.12 dB
nbcoeff $x_{\mathbf{U}} + x_{\mathbf{V}}$	23.8%	24.1	24.2%
SNR $x_{\mathbf{V}} + x_{\mathbf{V}}$	26.1 dB	25.4 dB	24.1 dB

Table 1 Results obtained for three different choices of estimates. number of retained coefficients in each layer and reconstruction, and corresponding SNR values. L stands for LASSO, G-L for G-LASSO and E-L for E-LASSO.

layer, defined as 20 times the base two logarithm of the ratio of the energy of the signal by the energy of the layer. The last two lines give the total number of retained coefficient and the SNR of the reconstruction $x_{\mathbf{U}} + x_{\mathbf{V}}$, defined as above. Here, SNRs should not be interpreted as a performance measure, but rather as a way to compare the behaviors of the three estimates. Hence, one can see that with the LASSO/LASSO choice, the transient layer is closer to the original signal than with the other two choices, but does not yield the best expected results for this layer (see figures and discussion below).

Together with this table, Fig. 11 and 12 clearly show the different behaviors of the estimators. Obviously, E-LASSO promotes persistence in comparison to LASSO. For the transient layer, the vertical structures are better preserved by E-LASSO. In comparison, the LASSO transient estimate catches a lot of low frequency components, which is generally not desirable.

The G-LASSO transient estimate (Fig. 12) performs quite differently. Like E-LASSO, it is not affected by low frequency components. In addition, it provides a very simple map of nonzero coefficients, which may be interesting for some tasks such as transient or onset detection. However, this estimate may also be considered an over simplification of the transient layer.

Even though the reconstructions obtained with the three decompositions are very similar (in terms of SNR and listening[2]), the behaviors of the layers are completely different. With the LASSO/LASSO choice, the low frequencies are shared between the two layers, while the partials are better preserved in the tonal layer with the LASSO/E-LASSO and LASSO/G-LASSO estimates. Fig. 11 shows the differences between the time-frequency coefficients for the three estimates of the tonal layer.

We also tried to replace E-LASSO estimate by its AE-LASSO approximation (17) (even though the convergence proof of Algorithm 1 is not valid any more in this case, we always observed numerical convergence).

[2] Soundfiles of the different estimates can be listened from the website [1].

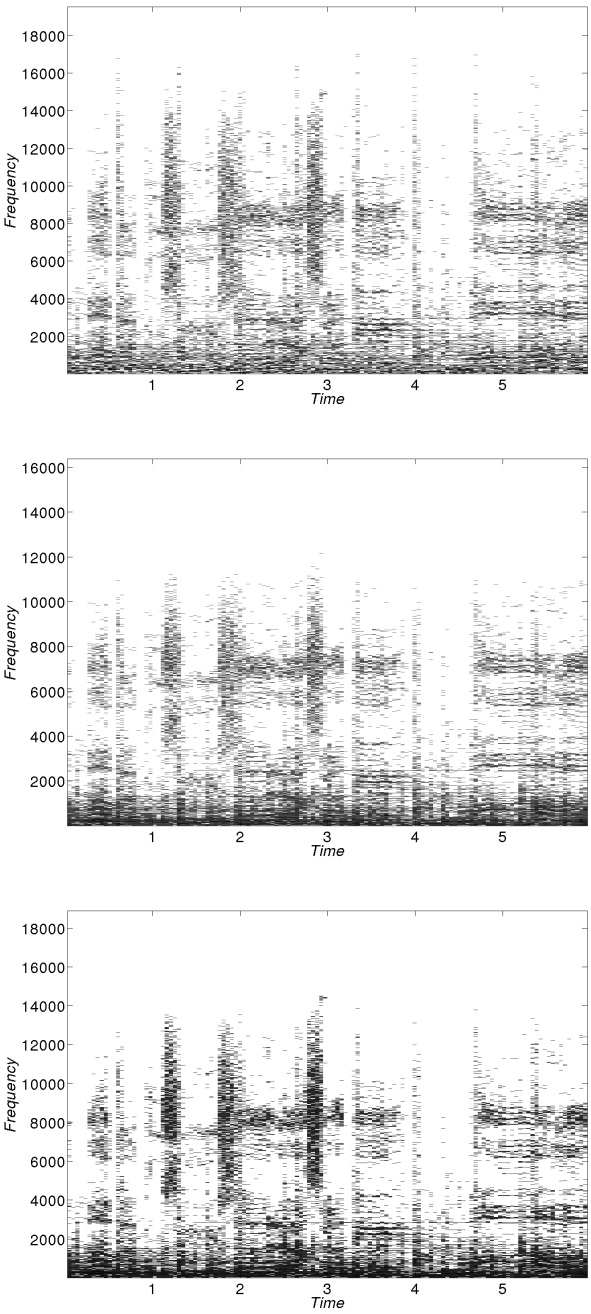

Fig. 11 MDCT coefficients of the three estimates of the tonal layer. From top to bottom: LASSO/LASSO, LASSO/E-LASSO, LASSO/G-LASSO

As suggested by the simulation of the previous section, results are similar to the E-LASSO, if one does not look for a very sparse estimate.

Let us stress that the main shortcoming of E-LASSO is the sensitivity to the regularization parameters λ and μ: slight changes to the parameters may affect the solution significantly. AE-LASSO appears to be much less sensitive to the choice of the regularization parameter.

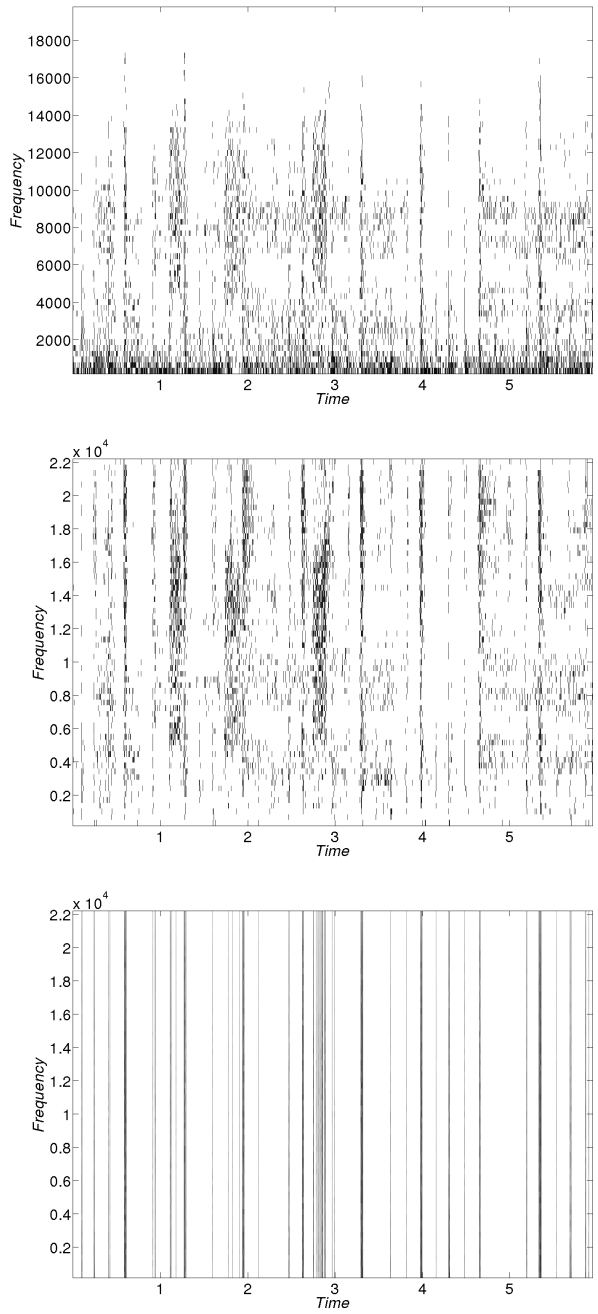

Fig. 12 MDCT coefficients of three estimates of the transient layer. From top to bottom: LASSO/LASSO, LASSO/E-LASSO, LASSO/G-LASSO.

5 Conclusion

We have shown in this paper the relevance of mixed norm priors in the framework of sparse regression problems. Such mixed norms have been extensively used in the mathematical analysis literature, but their use in practical situations are limited to some particular ones such as the G-LASSO and in the context of joint sparsity for multichannel signals. For the sake of simplicity, the mixed norms discussed here are the $\ell_{1,2}$ and $\ell_{2,1}$ norms, but similar results may be obtained using more general $\ell_{p,q}$ norms, and several standard sparse approximation algorithms may be extended to that situation. We refer to the forthcoming paper [14] for a thorough analysis of the latter.

Mixed norms yield generalised shrinkage operators; we also proposed new generalizations of the latter, that allow one to refine signal modelling, and overcome some shortcomings of standard thresholding operations. The E-LASSO estimate (and its approximation AE-LASSO) is a solution for the "over-sparsifying" behavior of the ℓ_1 norm. The WG-LASSO is a valuable alternative to G-LASSO when no well-defined group of coefficients is available. We applied these operators on simulated signal to illustrate as clearly as possible their respective behaviors.

Here, we have only emphasized a couple of applications, in the domain of audio signal processing, for which the results were encouraging. Let us stress that our point was not to compare to state of the art approaches, but rather to show what can be done using very simple techniques, that can be refined further. We would also like to point out that this approach is not at all specific to audio signals, and may be applied *mutatis mutandis* to image decomposition, for example in the framework of the MCA approach of [10], or multichannel signals such as EEG/MEG signals.

To conclude, it is worth coming back to the behavior of mixed norms in the present context. The rationale of our approach is to use a combination of ℓ_1 and ℓ_2 norms, to promote sparsity in the direction of one of the two indices, and persistence in the direction of the other. Now, as we have stressed at the beginning of this paper, a doubly labelled coefficient sequence can be obtained by arbitrary re-labelling of a given coefficient sequence. Therefore, mixed norm approaches can be used to introduce models for coefficients involving a small number of clusters of significant coefficients. Such a representation features both sparsity (in the domain of coefficient groups) and persistence (within a group). We believe that the potential of such approaches can be very important in a number of practical situations.

Acknowledgements

We wish to thank the anonymous referees for their comments, and for bringing a few additional references to our attention. We also thank Stéphane Molla for kindly providing the train sound example.

References

1. URL http://www.cmi.univ-mrs.fr/~kowalski/SIViP08.html
2. Berger, J., Coifman, R., Goldberg, M.: Removing noise from music using local trigonometric bases and wavelet packets. J. Audio Eng. Soc. **42**(10), 808–818 (1994)
3. Bobin, J., Moudden, Y., Fadili, J., Starck, J.L.: Morphological component analysis for sparse multichannel data: Application to inpainting (2007). Preprint, submitted
4. Bruce, A.G., Sardy, S., Tseng, P.: Block coordinate relaxation methods for nonparametric signal denoising. In: Proceedings of the SPIE - The International Society for Optical Engineering, 3391, pp. 75–86 (1998)
5. Chen, S.S., Donoho, D.L., Saunders, M.: Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing **20**(1), 33–61 (1998)
6. Cotter, S., Rao, B., Engan, K., Kreutz-Delgado, K.: Sparse solutions to linear inverse problems with multiple measurement vectors. IEEE Transactions on Signal Processing **53**(7), 2477–2488 (2005)
7. Daudet, L., Molla, S., Torrésani, B.: Towards a hybrid audio coder. In: J.P. Li (ed.) International Conference Wavelet analysis and Applications, pp. 13–24. Chongqing, China (2004)
8. Daudet, L., Torrésani, B.: Hybrid representations for audiophonic signal encoding. Signal Processing **82**(11), 1595–1617 (2002). Special issue on Image and Video Coding Beyond Standards
9. Donoho, D., Tsaig, Y., Drori, I., Starck, J.L.: Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. Tech. rep., Statistics Department, Stanford University (2007). Preprint
10. Elad, M., Starck, J.L., Donoho, D.L., Querre, P.: Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). Journal on Applied and Computational Harmonic Analysis **19**, 340–358 (2005)
11. Févotte, C., Torrésani, B., Daudet, L., Godsill, S.J.: Sparse linear regression with structured priors and application to denoising of musical audio. IEEE Transactions on Audio Speech and Language Processing **16**(1), 174–185 (2008)
12. Fornasier, M., Rauhut, H.: Recovery algorithm for vector-valued data with joint sparsity constraints. SIAM Journal on Numerical Analysis (to appear) **46**(2), 577–613 (2007)
13. Gribonval, R., Rauhut, H., Schnass, K., Vandergheynst, P.: Atoms of all channels, unite! average case analysis of multichannel sparse recovery using greedy algorithms (2007). INRIA technical report PI 1848, submitted
14. Kowalski, M.: Sparse regression using mixed norms (2008). URL http://hal.archives-ouvertes.fr/hal-00202904/
15. Molla, S., Torrésani, B.: An hybrid audio scheme using hidden Markov models of waveforms. Applied and Computational Harmonic Analysis **18**(2), 137–166 (2005)
16. Samarah, S., Obeidat, S., Salman, R.: A Shur test for weighted mixed-norm spaces. Analysis Mathematica **31**, 277–289 (2005)
17. Starck, J.L., Elad, M., Donoho, D.: Image decomposition via the combination of sparse representations and a variational approach. IEEE Transactions on Image Processing **14**(10) (2004)
18. Teschke, G., Ramlau, R.: An iterative algorithm for nonlinear inverse problems with joint sparsity constraints in vector valued regimes and an application to color image inpainting. Inverse Problems **23**(5), 1851–1870 (2007)
19. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Serie B **58**(1), 267–288 (1996)
20. Tropp, J., Gilbert, A., Strauss, M.: Algorithms for simultaneous sparse approximation. part I: Greedy pursuit. Signal Processing **86**, 572–588 (2006). Special issue "Sparse approximations in signal and image processing"
21. Tropp, J., Gilbert, A., Strauss, M.: Algorithms for simultaneous sparse approximation. part II: Convex relaxation. Signal Processing **86**, 589–602 (2006). Special issue "Sparse approximations in signal and image processing"
22. Tseng, P.: Dual coordinate ascent methods for non-stricly convex minimization. Mathematical Programming **59**, 231–247 (1993)
23. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Serie B **68**(1), 49–67 (2006)